DOCUMENT RESUME

ED 381 555 TM 022 890

AUTHOR Wainer, Howard; Thissen, David
TITLE On Examinee Choice in Educational Testing. GRE Board Professional Report No. 91-17P.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
REPORT NO ETS-RR-94-31
PUB DATE Jun 94
NOTE 42p.; Reprint from "Review of Educational Research," Spring 1994, Vol. 64, No. 1, pp 159-195.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Constructed Response; *Difficulty Level; *Educational Testing; Equated Scores; *Individual Testing; *Item Banks; Literature Reviews; *Test Bias; Test Construction; Test Format; Test Items; Test Length
IDENTIFIERS *Choice Behavior

ABSTRACT

When an examination consists in whole or part of constructed response test items, it is common practice to allow the examinee to choose a subset of the constructed response questions from a larger pool. It is sometimes argued that, if choice were not allowed, the limitations on domain coverage forced by the small number of items might unfairly affect some examinees. Alternatives, such as increasing test length or confining questions to a core curriculum, might discourage teachers because of practical considerations. In this consideration of whether allowing examinee choice is a sensible strategy, some of the pitfalls of allowing choice are described. Some experimental steps that can tell whether choice can be implemented fairly are discussed. A bleak picture of the use of examinee choice emerges. To make tests with choice fair requires equating the test forms generated by the tests for their differential difficulty. Accomplishing this requires some special data gathering effort or trust in assumptions about unobserved responses that, if true, obviate the need for choice. If test items can be equated successfully, the value of choice is removed for any but the most superficial sense. Eight tables and seven figures illustrate the discussion. (Contains 40 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 381 555

GRE
RESEARCH

On Examinee Choice in Educational Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Coley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Howard Wainer
and
David Thissen

June 1994

GRE Board Professional Report No. 91-17P
ETS Research Report 94-31

Educational Testing Service, Princeton, New Jersey

BEST COPY AVAILABLE

On Examinee Choice in Educational Testing

Howard Wainer
and
David Thissen

GRE Board Report No. 91-17P

June 1994

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1994 by the American Educational Research Association.
Reprinted by permission of the publisher.

Review of Educational Research
Spring 1994, Vol. 64, No. 1, pp. 159–195

On Examinee Choice in Educational Testing

Howard Wainer
Educational Testing Service
David Thissen
University of North Carolina at Chapel Hill

Throughout our society, individuals are compared on summaries of diverse measures. Often the measures are neither the same for all competitors nor selected randomly from a universe of measures. In fact, comparisons often are based on a mixture of measures in which the competitors choose some or all of the elements of the mixture; applicants to colleges choose some of the courses they take in high school, which extracurricular activities they participate in, and (to some extent) which entrance exams they take. What are the consequences of any selection procedure, if such choices are allowed? How can the adverse consequences be minimized? This article summarizes results from tests that have allowed examinee choice heretofore and provides a basis for a fuller discussion of these issues.

> If you allow choice, you will regret it; if you don't allow choice, you will regret it; whether you allow choice or not, you will regret both.
>
> —Kierkegaard, 1986, p. 24

Throughout the educational process, decisions are made using nonrandomly selected data. Admissions to college are decided among individuals whose dossiers contain mixtures of material; one high school student may opt to emphasize courses in math and science whereas another may have taken advanced courses in French and Spanish. One student might have been editor of the school newspaper, another captain of the football team, and yet a third might have been first violin in the orchestra. All represent commitment and success; how are they to be compared? Is your French better than my calculus? Is such a comparison sensible? Admissions offices at competitive universities face these problems all the time; dismissing them is being blind to reality. Moreover, there is obvious sense in statements like, "I know more physics than you know

This report summarizes the wisdom gathered during the course of research carried on over the past 3 years and owes part of its content to contributions from Paul Holland, Robert Lukhele, Donald Rubin, and Xiang-bo Wang. In addition, we have profited from conversations with Lyle Jones, Skip Livingston, Nick Longford, and Rick Morgan. We are pleased to thank them. Of course, none of these worthy scholars should be held responsible for the opinions expressed or any errors that might remain. Last, this report, and most of the research described herein, was supported in whole, or in part, by the Graduate Record Board. We gratefully acknowledge this support, without which little of this would have been accomplished.

French." Or, "I am a better runner than you are a swimmer." Cross-modal comparisons are not impossible, given that we have some implicit underlying notion of quality. How accurate are such comparisons? Can we make them at all when the differences between the individuals are subtle? How do we take into account the difficulty of the accomplishment? Is being an All-State athlete as distinguished an accomplishment as being a Merit Scholarship finalist?

How can we understand comparisons like these? Perhaps we can begin by considering the more manageable situation that manifests itself when the examination scores of students who are to be compared are obtained from test items that the students have chosen themselves. Such a situation is likely to occur increasingly often in the future, because of the greater contemporary emphasis on assessing what are called *generative* or *constructive processes* in learning. To be able to measure such processes, testing programs believe they must incorporate constructed response items into their previously multiple-choice standardized exams. It is hoped that large items such as testlets, essays, mathematical proofs, experiments, portfolios, or other performance-based tasks are better able to measure deep understanding, broad analysis, and higher levels of performance than traditional multiple-choice items. Examples of tests currently using large items are the College Board's Advanced Placement Examinations in United States History, European History, United States Government and Politics, Physics, Calculus, and Chemistry. There are many others.

When an exam consists, in whole or in part, of constructed response items, it is common practice to allow the examinee to choose a subset of the constructed response questions from a larger pool. It is sometimes argued that, if choice were not allowed, the limitations on domain coverage forced by the small number of items might unfairly affect some examinees. An alternative to choice would be to increase the length of the test; this is not often practical. Another alternative would be to confine the test questions to a core curriculum that all valid courses ought to cover. This option may discourage teachers from broadening their courses beyond that core.

Even in the absence of attractive alternatives, is allowing examinee choice a sensible strategy? Under what conditions can we allow choice without compromising the fairness and quality of the test? We will not provide a complete answer to these questions. We will illustrate some of the pitfalls associated with allowing examinee choice; we will provide a vocabulary and framework that aids in the clear discussion of the topic; we will outline some experimental steps that can tell us whether choice can be implemented fairly; and we will provide some experimental evidence that illuminates the topic. We begin with a brief history of examinee choice in testing, in the belief that it is easier to learn from the experiences of our clever predecessors than to try to relive those experiences.

A Selective History of Choice in Exams

We shall confine our attention to some college entrance exams used during the first half of the 20th century in the United States. The College Entrance Examination Board (CEEB) began testing prospective college students at the turn of the century. By 1905, exams were offered in 13 subjects: English, French, German, Greek, Latin, Spanish, mathematics, botany, chemistry, physics, drawing, geography, and history (College Entrance Examination Board, 1905). Most

of these exams contained some degree of choice. In addition, all of the science exams used portfolio assessment, as it might be called in modern terminology. For example, on the 1905 Botany exam 37% of the grade was based on the examinee's laboratory notebook. The remaining 63% was based on a 10-item exam. The examinee was asked to answer 7 of those 10 items. This yielded 120 (10 choose 7) different examinee-created "forms."

Question 10 (below), of the 1905 Botany exam, could complicate the computation of the number of choice-created forms:

> *10. Select some botanical topic not included in the questions above, and write a brief exposition of it.*

The graders for each test were identified, with their institutional affiliations. The chemistry and physics exams shared the structure of the botany exam.

As the CEEB grew more experienced, the structure of its tests changed. The portfolio aspect disappeared by 1913, when the requirement of a teacher's certification that the student had, in fact, completed a lab course was substituted for the student's notebook. By 1921, this certification was no longer required.

The extent to which examinees were permitted choice varied; see Table 1, in which the number of possible examinee-created forms are listed for several subjects and years. The contrast between the flamboyance of the English exam and the staid German exam is instructive. Only once, in 1925, was any choice allowed on the German exam ("Answer only *one* of the following six questions"). The lack of choice seen in the German exam is representative of all of the foreign language exams except Latin. The amount of choice seen in the Physics and Chemistry exams parallels that for most exams that allowed choice. The English exam between 1913 and 1925 is unique in terms of the possible variation.[1] No equating of these examinee-constructed forms was considered.

By 1941, the CEEB offered 14 exams, but only 3 (American History, Contemporary Civilization, and Latin) allowed examinee choice. Even among these, choice was sharply limited:

- In the American History exam, there were six essay questions. Essays 1, 2, and 6 were mandatory. There were three questions about the American

TABLE 1
Number of possible test forms generated by examinee choice patterns

	Subject			
Year	Chemistry	Physics	English	German
1905	54	81	64	1
1909	18	108	60	1
1913	8	144	7,260	1
1917	252	1,620	1,587,600	1
1921	252	216	2,960,100	1
1925	126	56	48	6
1929	20	56	90	1
1933	20	10	24	1
1937	15	2	1	1
1941	1	1	1	1

Constitution (labeled 3A, 4A, and 5A) as well as three parallel questions about the British Constitution (3B, 4B, and 5B). The examinee could choose either the A questions or the B questions.

- In the Contemporary Civilization exam, there were six essay questions. Questions 1–4 and 6 were all mandatory. Question 5 consisted of six short-answer items out of which the examinee had to answer five.
- The Latin exam had many parts. In sections requiring translation from Latin to English or vice versa, the examinee often had the opportunity to pick one passage from a pair to translate.

In 1942, the last year of the program, there were fewer exams given; none allowed examinee choice. Why did the use of choice disappear over the 40 years of this pioneering examination program? Our investigations did not yield a definitive answer, although there are many hints that suggest that issues of fairness propelled the CEEB toward the test structure on which they eventually settled. Our insight into this was sharpened during the reading of Brigham's (1934, p. i) remarkable report on "the first major attack on the problem of grading the written examination." This exam required the writing of four essays. There were six topics; Topics 1 and 6 were required of all examinees, and there was a choice offered between Topics 2 and 3 and between Topics 4 and 5. The practice of allowing examinee choice was termed "alternative questions" (p. i).

Brigham noted (1934, p. 7) that,

> When alternative questions are used, different examinations are in fact set for the various groups electing the different patterns. The total score reported to the college is to a certain extent a sum of the separate elements, and the manner in which the elements combine depends on their intercorrelation. This subject is too complex to be investigated with this material.

Brigham's judgment of the difficulty is mirrored by Harold Gulliksen (1950, p. 338), who wrote that, "In general it is impossible to determine the appropriate adjustment without an inordinate amount of effort. Alternative questions should always be avoided."

Both of these comments suggest that, while adjusting for the effects of examinee choice is possible, it is too difficult to do within an operational context. We suspect that even these careful researchers underestimated the difficulty of satisfactorily accomplishing such an adjustment; their warnings are strongly reminiscent of Fermat's marginal comments about his only just proved theorem. This view was supported by Ledyard Tucker (personal communication, February 10, 1993), who said that, "I don't think that they knew how to deal with choice then. I'm not sure we know how now."

The core of the problem of choice is that, when it is allowed, the examinees' choices generate what can be thought of as different forms of the test. These forms may not be of equal difficulty. When different forms are administered, standards of good testing practice require that those forms be statistically equated so that individuals who took different forms can be compared fairly. Perhaps, through pretesting and careful test construction, it may be possible to make the differences in form difficulty sufficiently small that further equating is not required. As we will show, an unbiased estimate of the difficulty of any item can only be obtained from a random sample of the examinee population. The

CEEB made no attempt to get such a sample. Perhaps that is why Brigham said that, "This subject is too complex to be investigated with this material."

Are Choice Items of Equal Difficulty?

We have not yet found a situation in which it was plausible to believe that two choice questions were of equal difficulty. However, in only one case could this be shown unequivocally; in operational tests, there are always alternative explanations. These alternatives, although often unlikely, can never be dismissed entirely. To illustrate, we will give three examples and provide references to more. We will then describe the only circumstance we know of in which the difficulty of choice items was established. This last case will provide motivation for the subsequent discussion of how choice can be permitted and tests can still be fair.

Example 1: 1968 AP Chemistry Test. Shown in Table 2 is a result reported by Fremer, Jackson, and McPeek (1968) for a form of the College Board's Advanced Placement Chemistry Test. One group of examinees chose to take Problem 4, and a second group chose Problem 5. While their scores on the common multiple-choice section were about the same (11.7 vs. 11.2, out of a possible 25), their scores on the choice problem were very different (8.2 vs. 2.7, on a 10-point scale). There are several possible conclusions to be drawn from this; four among them are:

1. Problem 5 is a good deal more difficult than Problem 4.
2. Small differences in performance on the multiple-choice section translate into much larger differences on the free response questions.
3. The proficiency required to do the two problems is not strongly related to that required to do well on the multiple-choice section.
4. Item 5 is selected by those who are less likely to do well on it.

Investigation of the content of the questions, as well as studies of the dimensionality of the entire test, suggests that Conclusion 1 is the most credible. This interpretation would suggest that scores on these two examinee-created test forms ought to have been equated. They were not.

Example 2: 1989 AP Chemistry Test. In the 1989 version of the AP Chemistry Test much had changed, but there was still examinee choice. Figure 1 shows the expected score on each of two choice items (examinees could opt to answer just one of these two) as a function of examinee proficiency as measured by the entire test (Wainer, Wang, & Thissen, 1991). To calculate these trace lines, an untestable assumption about unobserved examinee performance had to be made; this assumption will be discussed in detail later. It will suffice for now to note that, even with the assumptions that usually underlie item response theory, we also

TABLE 2
Average scores on AP Chemistry 1968

Choice group	MC section	Choice problem	
		4	5
1	11.7	8.2	
2	11.2		2.7

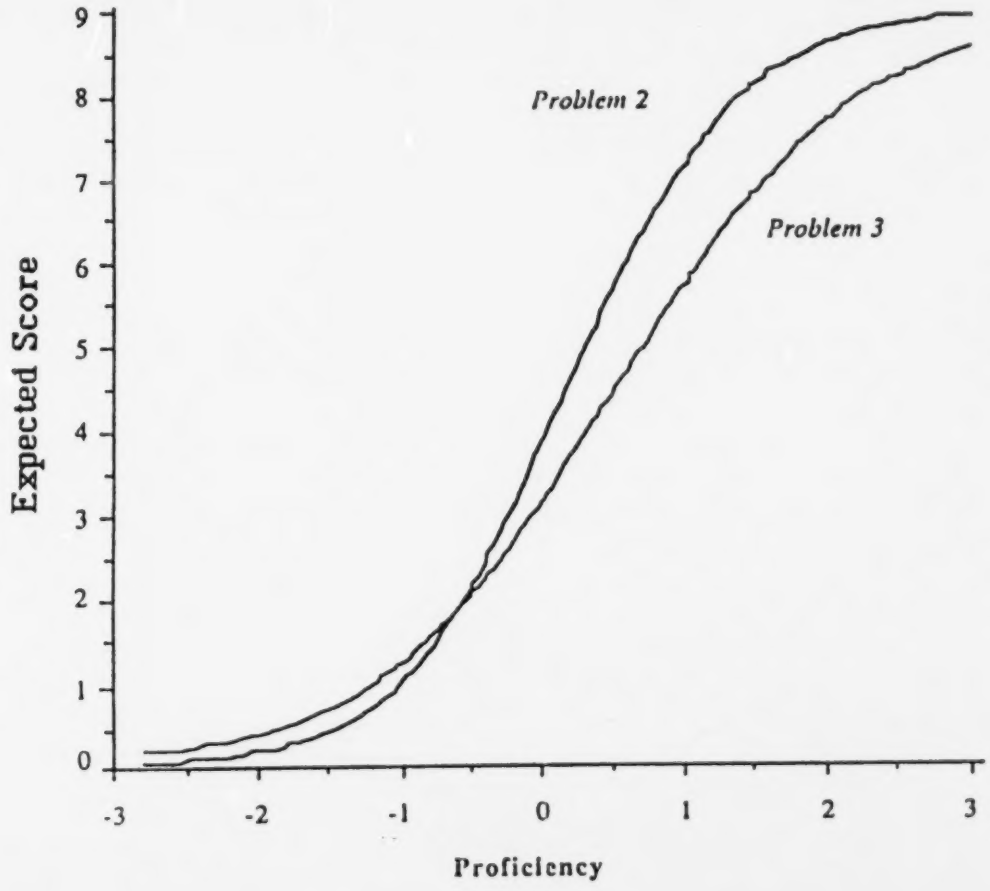

FIGURE 1. *A comparison of the expected score trace lines for the two choice problems in Part II, Section B, of the 1989 AP Examination in Chemistry*

had to assume that relatively poorer performance on Problem 3 by an examinee who otherwise performed equivalently to an examinee who chose Problem 2 must be reflected in differential difficulty of the choice problems, not multidimensionality.

There is about a one-point advantage for an examinee who takes Problem 2 and whose proficiency lies between 0 and 2 on this scale. This is an improvement on the 1968 test, but it is by no means a trivial difference. To put this in perspective, there are about 160 points on the test, and a score of about 45 is sufficient to obtain college credit for completing a one-semester chemistry course. One point is not a trivial amount.

Example 3: 1988 AP United States History Test. The first two examples indicated that those who chose the more difficult problems were placed at a disadvantage. In this example, we identify more specifically the examinees who tended to choose more difficult items. Consider the results shown in Figure 2 from the 1988 administration of the College Board's Advanced Placement Test in United States History (hereafter AP US History). This test comprises 100

multiple-choice items and two essays. The first essay is mandatory; the second is chosen by the examinee from among five topics. Shown in Figure 2 are the average scores given for each of those topics, as well as the proportion of men and women who chose them.

Essay 3 had the lowest average scores for both men and women. The usual interpretation of this finding is that the topic of Essay 3 was the most difficult. This topic was about twice as popular among women as among men. An alternative interpretation of these findings might be that the lowest proficiency examinees chose this topic and that a greater proportion of women than men fell into this category. This illustrates again how any finding within a self-selected sample yields ambiguous interpretations.

This same phenomenon shows itself on the chemistry test as well. Shown in Table 3 are the summary proportions of male and female examinees choosing various triples in a choose-3-of-5 section of the 1989 AP Chemistry Test. Of the five choice items, Item 6 is the most difficult, and Items 8 and 9 are the easiest. Note that female examinees chose Item triples 5,6,7 and 5,6,8 considerably more often than male examinees; men chose the easier items more often than women. Because all items counted equally, men had an advantage: not because of any superior knowledge of chemistry but merely because they tended to choose easier items to answer.

Similar studies with similar findings have been carried out on all of the AP tests that allow choice (DeMauro, 1991; Pomplun, Morgan, & Nellikunnel, 1992). Fitzpatrick and Yen (1993) also reported substantial sex and ethnic differences in

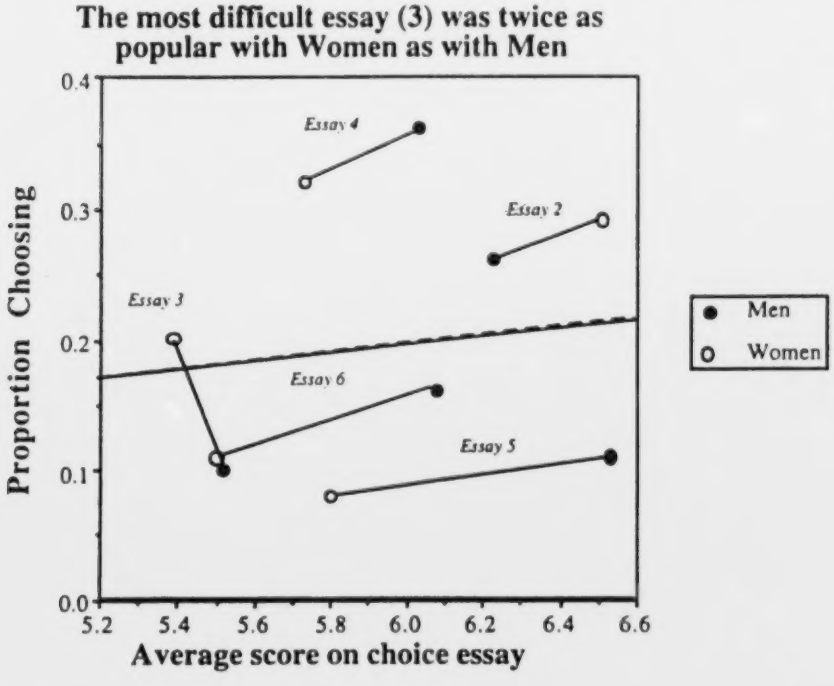

FIGURE 2. *The relative performance of men and women on the choice essays in the 1988 AP Examination in History*

TABLE 3
Item 6 is the hardest; Items 8 and 9 are the easiest

Choice on Section D	Proportion choosing		
	Males	Females	
567	0.13	0.16	Hardest
568	0.25	0.35	
578	0.28	0.24	
789	0.06	0.03	
589	0.10	0.07	Easiest

choice on third-, fifth-, and eighth-grade passage-based reading comprehension tests. In the data described by Fitzpatrick and Yen, there is no obvious tendency of any particular group to select more or less difficult items; however, for any particular form of the test, the outcome of the combination of choice and scoring was that some group was placed at a disadvantage.

The phenomenon of students choosing poorly is so widespread that evidence of it occurs whenever one looks for it. It is instructive to note that Powers, Fowles, Farnum, and Gerritz (1992), in their description of the effect of choice on a test of basic writing, found that the more that examinees liked a particular topic, the lower they scored on an essay they subsequently wrote on that topic. Their results are summarized in Figure 3. An examinee's preference for a topic does not predict how well he or she will do on it. Perhaps if examinees were explicitly informed that they ought to choose a topic on the basis of how well they think they can score, and not how much they like the topic, they might choose more wisely.

Although test developers try to make up choice questions that are of equivalent difficulty, they do not appear to have been completely successful. Of course, this conclusion is clouded by possible alternative explanations that have their origin in self-selection: For example, the choice items are not unequally difficult; rather, the people who chose them are unequally proficient. So long as there is self-selection, this alternative cannot be completely dismissed, although it can be discredited through the use of covariate information. If it can be shown that choice items are not of equal difficulty, it follows that some individuals will be placed at a disadvantage by their choice of item—they choose to take a test some of the items of which are more difficult than those on a corresponding test for other examinees—and this extra difficulty is not adjusted away in the scoring.

The only unambiguous data on choice and difficulty that we know of involved data gathered and reported by Wang, Wainer, and Thissen (1993) in which examinees were presented with a choice of two items but then required to answer both. In Table 4, we see that, even though Item 12 was much more difficult than Item 11, there were still some students who chose it. The item difficulties in Table 3 were obtained from an operational administration of these items involving more than 18 thousand examinees. Perhaps examinees chose Item 12 because they had some special knowledge that made this item less difficult for them than Item 11. In Table 5 is shown the performance of examinees on each of these items broken down by the items they chose. Note that 11% of those examinees who

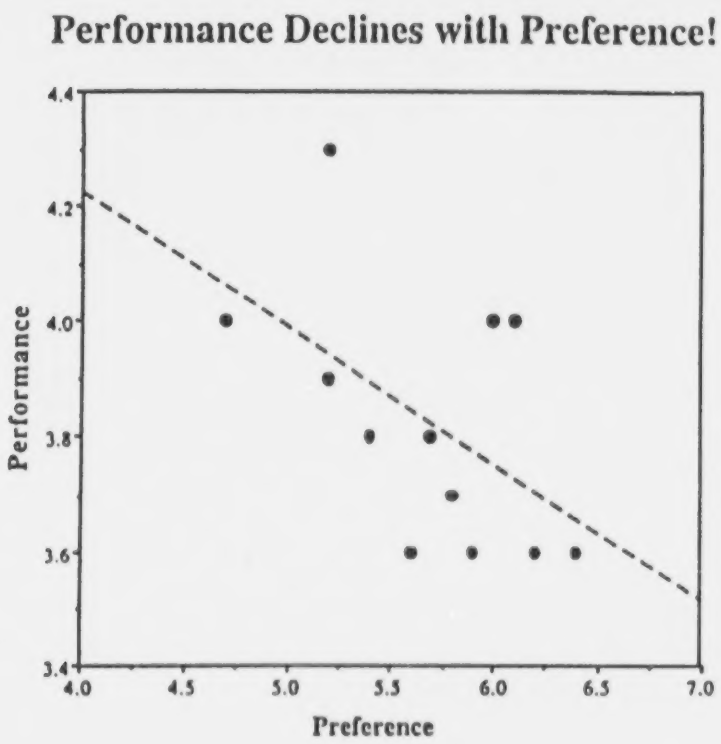

FIGURE 3. *In a test of basic writing, examinees scored lower on essay topics that they liked more*

Note. Adapted from Powers, Fowles, Farnum, and Gerritz, 1992, Table 3, p. 17.

chose Item 12 responded correctly, whereas 69% of them answered Item 11 correctly. Moreover, this group performed more poorly on both of the items than examinees who chose Item 11. The obvious implication drawn from this example is that examinees do not always choose wisely and that less proficient examinees exacerbate matters through their unfortunate choices. Data from the rest of this experiment consistently supported these conclusions.

Strategies for Fair Testing When Choice Is Allowed

We will now define more carefully the goal of examinee choice. We will then examine two alternative strategies for achieving that goal. The first is to aid

TABLE 4

The number of students choosing each item and the difficulty of those items in the general test-taking population

Item chosen	Number choosing	Item difficulty (*b*)
11	180	−2.5
12	45	1.6

TABLE 5
The proportion of students getting each item correct shown conditional on which item they preferred to answer

Item answered	Item chosen	
	11	12
11	0.84	0.69
12	0.23	0.11

students in making wiser choices; the second is to diminish the unintended consequences of a poor choice through statistical equating of the choice items.

What Do We Get When We Allow Choice?

When a test is given we ordinarily estimate a score. Following traditional IRT notation,[2] we will call the proficiency of a particular examinee taking the test θ and the estimate of that proficiency $\hat{\theta}$. Ordinarily, each item on a test provides a somewhat different value of the examinee's estimated proficiency; sometimes the variance of the distribution of those estimates is used as an index of measurement precision (Wainer & Wright, 1980).

θ_{Max}. The usual estimate of proficiency is based on estimated performance from a random sample drawn from the entire distribution of items. But suppose we allow choice. What are we estimating? Most practitioners we have spoken to, who favor allowing choice, argue that choice provides the opportunity for the examinees to show themselves to best advantage. We shall call this proficiency θ_{Max}.

Figure 4 is a graphical depiction of what might occur in a choice situation. The top panel is the distribution of θ based on information obtained prior to administering the choice items. The distribution is centered on 0. Beneath this panel is a second panel showing the trace lines for two items, assuming that they are answered correctly. Item 1 is relatively easy; Item 2 is somewhat more difficult. Each curve in the bottom panel is constructed by multiplying the prior density by the appropriate trace line (or one minus that trace line, for incorrect responses). These are called the posterior densities after choice and represent our knowledge of that examinee's ability.

θ_{Max} is the characterization of proficiency that would be obtained if the examinees chose the item that would give them the highest score. When we allow choice we are attempting to estimate θ_{Max}. What we are actually obtaining is $\hat{\theta}_{Max}$. How is $\hat{\theta}_{Max}$ related to θ_{Max}?

To answer this question, it is useful to adopt a Bayesian approach. But first, we must specify the scoring system for the items, because an examinee's expected score is $P_{\text{item}}s$, where P_{item} is the probability the examinee will obtain score s from the item. We will consider two possibilities:

1. Both items are scored $s = 1$, if correct, $s = 0$, if incorrect (simple summed scoring). In this case, only the information in the second panel of Figure 4 is relevant to the choice: We see that $P_1 > P_2$ for all values of θ, so all examinees maximize their expected score by choosing Item 1.

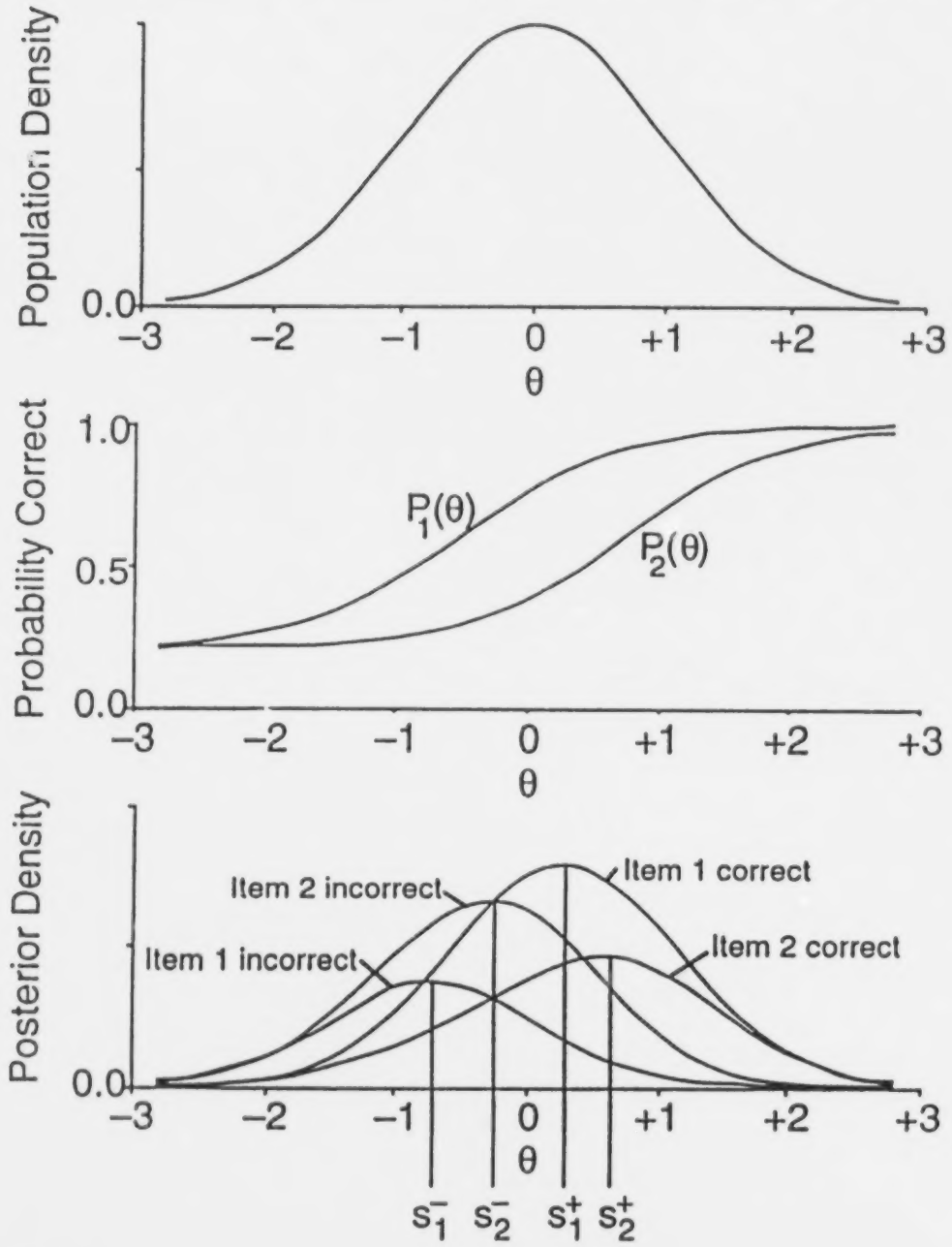

FIGURE 4. *A graphical representation depicts how the posterior density is computed from the product of the prior density and the appropriate item trace lines*

2. The items are scored using IRT—for example, using the mean of the posterior given the item response. In this case, the score associated with Item 1 correct is s_1^+; that for Item 1 incorrect is s_1^-. The score associated with Item 2 correct is s_2^+; that for Item 2 incorrect is s_2^-.[3] The locations of the

scores are shown in Figure 4. Now the examinee's task is more difficult; the choice must be Item 2 if and only if

$$P_2 s_2^+ + (1 - P_2)s_2^- > P_1 s_1^+ + (1 - P_1)s_1^-.$$

To do this computation, examinees must know both P_1 and P_2 (at least for their own values of θ), as well as the item scores.

But the examinees, when choosing an item, do not know their probability of responding correctly to the item. Instead, they have some subjective idea of that probability. We have already provided strong evidence indicating that some examinees do not choose wisely. Moreover, we have seen that the propensity for making optimizing choices varies by sex and ethnic group.

As we have already seen, choice items, as currently prepared, are typically not of equal difficulty. This fact, combined with the common practice of not equating choice items for their differential difficulty, yields the inescapable conclusion that it matters what choice an examinee makes. Examinees who chose the more difficult question will, on average, get lower scores than would have been the case had they chosen the easier item. The fact that all examinees do not choose those items that will show their proficiency to best advantage completes this unhappy syllogism: examinee choice is not likely to yield credible estimates of θ_{Max}.

What Can We Do to Improve Matters?

There appear to be two paths that can be followed: eliciting wiser choices by examinees or equating test forms. The second option removes the necessity for the first; in fact, it makes examinee choice unnecessary.

How can we improve examinees' judgment about which items to select? Estimation of θ_{Max} can be done optimally only by asking examinees to answer all items and then scoring just those responses that yield the highest estimate of performance. This strategy is not without its drawbacks. First, it takes more testing time, and choice is often instituted to keep testing time within practical limits. Second, many examinees, on hearing that "only one of the six items will be counted" will only answer one. Thus, this strategy may commingle measures of grit, choice wisdom, and risk aversion with those of proficiency.

A more practical approach might be to try to improve the instructions to the examinees about how the test is graded, to guide their choices better. It would be well if the instructions about choice made it clear that there is no advantage to answering a hard item correctly relative to answering an easy one, if such is indeed the case. Current instructions do not address this issue. For example, the instructions about choice on the 1989 AP Chemistry Test (CEEB, 1990, p. 23), reproduced in their entirety are:

> Solve ONE of the two problems in this part. (A second problem will not be scored.)

Contrast this with the care that is taken to instruct examinees about the hazards of guessing. These are taken from the same test (p. 3):

> Many candidates wonder whether or not to guess the answers to questions about which they are not certain. In this section of the examination, as a correction for haphazard guessing, one-fourth of the number of questions you answer incorrectly will be subtracted from the number you answer

> correctly. It is improbable, therefore, that mere guessing will improve your score significantly; it may even lower your score, and it does take time. If, however, you are not sure of the correct answer but have some knowledge of the question and are able to eliminate one or more of the answer choices as wrong, your chance of getting the right answer is improved, and it may be to your advantage to answer such a question.

Perhaps, with better instructions, the quality of examinee choices can be improved. At the moment, there is no evidence supporting the conjecture that they can be, or if so, by how much. An experimental test of the value of improved instructions could involve one randomly selected group with the traditional instructions and another with a more informative set; which group has higher average scores on the choice section? A more complex experiment could use a paradigm much like that employed by Wang, Wainer, and Thissen (1993) in which examinees were asked to choose from among several items but were then required to answer all of them. This sort of experiment would allow a detailed examination of the change in choice behavior due to the instructions. While more explicit instructions may help matters somewhat, and ought to be included regardless of their efficacy, we are not sanguine about this option solving the problem of getting $\hat{\theta}_{Max}$ closer to θ_{Max}. To do this requires reducing the impact of unwise choice.

As we pointed out, there are two terms in the calculation of subjective posterior density. The first is the subjective probability of getting the item correct; improved instructions may help this. The second requires that the examinee have an accurate idea of the relative difficulty of the choice items. Pretesting, when it is possible, would allow us to present to examinees each item's difficulty in the pretest population. It will not help to characterize the individual variations in item difficulty that are the principal reason for allowing choice. A more promising path seems to be to make all of the choice problems equally difficult (from the point of view of the entire examinee population) and allow the choice to be governed by whatever special knowledge or proficiency each individual examinee might possess. In this way, we can be sure that, at least on average, the items are as fair as possible. The problem is that it is at least difficult, and perhaps impossible, to build items that empirically turn out to be exactly equal in difficulty. Another option is to adjust the scores on the choice items statistically for their differential difficulty. We will refer to this statistical adjustment as *equating*, although the way it is carried out may not satisfy the strict rules that are sometimes associated with that term.

How Does Equating Affect the Examinee's Task?

Equating appears, at first blush, to make the examinee's task of choosing more difficult still. If no equating is done, the instructions to the examinee should be:

> Answer that item that seems easiest to you

(and we hope that the examinees choose correctly, but we will not know if they do not). If we equate the choice items (give more credit for harder items than easier ones), the instructions should be:

> Pick that item which, after we adjust, will give you the highest score.

This task could be akin to the problem faced by competitive divers, who choose their routine of dives from within various homogeneous groups of dives. The diver's decision is informed by:

- knowledge of the degree of difficulty of each dive,
- knowledge of the concatenation rule by which the dive's difficulty and the diver's performance rating are combined (they are multiplied), and
- knowledge, obtained through long practice, of what his or her score is likely to be on all of the dives.

Armed with this knowledge, the diver can select a set of dives that is most likely to maximize his or her total score.

The diver scenario is one in which an individual's informed choice provides a $\hat{\theta}_{Max}$ that seems to be close enough to θ_{Max} for useful purposes. Is a similar scenario possible within the plausible confines of standardized testing? Let us examine the aspects of required knowledge point by point.

Specifying how much each item will count in advance is possible, either by calculating the empirical characteristics of each item from pretest data or, as is currently the case, by specifying how much each one counts by fiat. We favor the former, because it allows each item to contribute to total score in a way that minimizes measurement error. An improvident choice of a priori weights can have a serious deleterious effect on measurement accuracy (see Lukhele, Thissen, & Wainer, 1993; Wainer & Thissen, 1993a).

Specifying the concatenation rule (how examinee performance and item characteristics interact to contribute to the examinee's score) in advance is also possible but may be quite complex, for example, if IRT is used. Perhaps a rough approximation can be worked out, or perhaps one could present a graphical solution like that shown in Figure 4, but for now this remains a question. The difficulties that we might have with specifying the concatenation rule are largely technical, and workable solutions could probably be developed.

A much more formidable obstacle is providing the examinees with enough information so that they can make wise choices. This seems completely out of reach, for, even if examinees know how much a particular item will, if answered correctly, contribute to their final score, it does no good unless the examinees have a good idea of their likelihood of answering the item correctly. The extent to which such knowledge is imperfect would then correspond to the bias (used in its statistical sense) associated with the use of $\hat{\theta}_{Max}$ to estimate θ_{Max}. The nature of security associated with modern large-scale tests makes impossible the sort of rehearsal that provides divers with accurate estimates of their performance under various choice options.

The prospect appears bleak for simultaneously allowing choice and satisfying the canons of good practice that require the equating of test forms of unequal difficulty. The task that examinees face in choosing items when they are adjusted seems too difficult. But is it? There remain two glimmers of hope. The brighter of these rests on the possibility of successfully equating the various choice forms. If we can do this, the examinees should be indifferent as to which items they answer, because successful equating means that an examinee will receive, in expectation, the same score regardless of the form administered. This is happy

but ironic news, for it appears that we can allow choice and have fair tests only when choice is unnecessary.

A dimmer possibility is to try to improve examinees' estimates of their success on the various choices. In a computer administered test, it may be possible to provide some model-based estimates of an examinee's probable score on each item. To the extent that these estimates are accurate, they might help. Of course, if really good estimates were available, we would not need to test further. Moreover, the value of choice would be greatest when an examinee's likelihood of success is very different than that predicted from the rest of the test.

To answer the question posed at the beginning of this section: When we do not equate selected items, the problem of choice faced by the examinee can be both difficult and important. When we do equate, the selection problem simultaneously becomes much more difficult but considerably less important. This conclusion naturally brings us to the next question.

Under What Conditions Can We Equate Choice Items? How?

Let us reconsider Harold Gulliksen's (1950, p. 338) advice, "Alternative questions should always be avoided." We have discussed one possible reason for this—that it makes the examinee's task too difficult. Our conclusion was that, while it does make the task difficult, this difficulty becomes irrelevant for most uses of the test score if the alternate forms thus constructed can be equated. This raises a second possible explanation for this advice: The equating task is too difficult. Certainly this explanation was the one favored by Tucker (quoted earlier). The only way to equate test forms that are created by choice is to make some (untestable) assumptions about the structure of the missing data that have resulted from the choice behavior.

One possible assumption is *missing-completely-at-random*. Underlying this assumption is the notion that, if we had the examinee's responses to all of the items, a random deletion of some portion of them would yield, in expectation, the same score as was obtained through the examinee's choice. In simple terms, we assume that the choice had no effect on the examinee's score. If we really believed missing-completely-at-random, we could equate without any anchor items because an important consequence of missing-completely-at-random is that all choice groups will have the same proficiency distribution. Data gathered from all of the Advanced Placement exams (Pomplun, Morgan, & Nellikunnel, 1992) suggest that this is not credible.

Thus, it is imperative to use required anchor items to establish a common scale for the choice items. This can be done using traditional or IRT methods (see Dorans, 1990) and is justified if we believe that the missing responses yielded by examinee choice are generated by a process that is, in Little and Rubin's (1987) terminology, conditionally ignorable.[4] What we mean by this weaker assumption is that the probability of an examinee choosing any particular item is independent of his or her likelihood of getting that item correct, conditional on θ. In graphical terms, this means that an item's trace lines are the same for those individuals who chose it as they would have been for those who omitted it.

Subsequent discussions will be clearer if we repeat our characterization of the missing data assumption and the logic surrounding their genesis with more precision.

Therefore, suppose

y_i is the score on test item Y_i, and
R_i is a choice function that takes the value 1 if Y_i is chosen and 0 if not.

In a choice situation, we can observe the distribution of scores, $f_1(y)$, for those who opted to take an item. This can be denoted

$$f_1(y) = P(Y = y | R = 1).$$

What we do not know, but what is crucial if we are to be able to equate the different choice items, is the distribution of scores, $f_0(y)$, for those who did not take the item. This is denoted

$$f_0(y) = P(Y = y | R = 0).$$

To be able to equate, we need to know the distribution of scores in the unselected population, $g(y) = P(Y = y)$.

Note that we can represent $g(y)$ as

$$g(y) = f_1(y) \times P(R = 1) + f_0(y) \times P(R = 0). \quad (1)$$

The only piece of this which is unknown is $f_0(y)$, the distribution of scores among those individuals who chose not to answer it. Unless one engages in a special data gathering effort, in which those examinees who did not answer Y_i are forced to, $f_0(y)$ is not only unknown but unknowable. Thus, the conundrum is that we must equate to ensure fairness, but we cannot equate without knowing $f_0(y)$.

One approach to such problems, mixture modeling (Glynn, Laird, & Rubin, 1986), involves a hypothesized structure for $f_0(y)$. It is convenient to assume that the function $f_0(y)$ is the same as $f_1(y)$. In formal terms,

$$\begin{aligned} f_1(y) &= P(Y = y | R = 1, \theta) = P(Y = y | R = 0, 0) = f_0(y) \\ &= P(Y = y | \theta). \end{aligned} \quad (2)$$

Or: We assume that the trace lines for the choice item would have been the same for those who didn't choose it as it was for those who did. If we could gather the appropriate data (forcing those who opted not to answer it to do so), this hypothesis could easily be tested using standard DIF technology (Holland & Wainer, 1993).

Although the conditional independence, given R and θ, expressed in Equation 2 has a surface similarity to the conditional independence, given θ, that underlies all of IRT, Equation 2 expresses a much stronger assumption that may or may not be true: Equation 2 states that, if θ is known, knowledge of whether the examinee chooses to answer an item or not does not affect the modeled probability of each response. This assumption is certainly contrary to the perceptions of examinees, who often feel that their choice of an item optimizes their score. However, there is little evidence available that illuminates the relationship between examinees' preference for a particular item and their eventual score. Contrary to widespread belief, what little experimental evidence there is supports the assumption expressed in Equation 2.

Thus, in the absence of contrary data, and because this assumption allows us to employ the existing technology of IRT to equate, we shall use it. For a fuller

description of the structure and consequences of assumptions about missing data, the reader is referred to Allen and Holland (1993); of special importance in the examinee choice situation is their distinction between ignorable and forgettable nonresponse.

What Other Assumptions Are Necessary for Equating?

While ignorable nonresponse is the only assumption that is new to this circumstance, it is not the only assumption required. In addition, we need to assume unidimensionality and fit to the test scoring model employed. These two latter assumptions are well known and can be tested with the test data ordinarily gathered; ignorable nonresponse cannot be. To test ignorable nonresponse requires a special data gathering effort. One example is the sort of data gathering scheme that Wang, Wainer, and Thissen (1993) employed: asking examinees to choose items but then requiring them to answer some of the items they did not choose. This is called *sampling from the unselected population* and will be discussed in greater detail later.

Equating test forms constructed by examinee choice can be straightforward once we have made some assumptions about the unobserved distribution of scores $f_0(y)$. While one can derive a formal equating procedure for many assumed characterizations of the missing data, the assumption of conditionally ignorable nonresponse allows us to immediately use the existing machinery for IRT equating. One merely enters the various vectors of item responses and treats what's missing as having not been presented to the individual. We establish a common scale by requiring a subset of items that all examinees must answer. This anchor test provides a set of items drawn from the unselected population on which we can also test model fit and unidimensionality.

How Can We Test Our Assumptions?

The special assumption required to equate choice items involves the distribution of scores on the choice items from those who did not answer them: $f_0(y)$. This distribution is necessary to estimate $g(y)$, the distribution of scores in the unselected population. There are many ways to test the viability of this assumption, but they all require some sort of special data gathering. We will describe two experimental designs that can be used to accomplish this.

Design 1: Within subjects. In a randomly chosen subset of the examinee population, examinees must be required to indicate their choice but then required to answer all items. This design allows us to estimate all three parts of Equation 1 and so allows an explicit test of the assumption stated as Equation 2. A variant of this design was employed by Wang (1992) which asked examinees their choices both before and after they answered the questions.

This design is subject to the criticism that examinees might not be particularly judicious in their choices when they know that they will have to answer all the questions anyway. If this conjecture is true, it is likely to affect the estimates of $f_0(y)$ and $f_1(y)$ more than that of the composite $g(y)$. Using the good estimates of $f_1(y)$ we can get from the operational choice test and the estimates of $g(y)$ from the experimental administration, we can derive $f_0(y)$ through Equation 1.

We used this design to test the assumption of ignorable nonresponse among some choice items in the 1989 AP Chemistry Test (Wang, Wainer, & Thissen,

1993), using IRT-based DIF technology (Thissen, Steinberg, & Wainer, 1988, 1993; Wainer, Sireci, & Thissen, 1991).

Figure 5 shows the estimated trace lines for choice Items 11 and 12 for those examinees that chose each item [$f_1(y)$] as well as for those that did not [$f_0(y)$]. The apparent difference between the two trace lines for Item 11 is somewhat unlikely ($\chi^2_{(2)} = 4$), whereas there is no difference at all between the two trace lines for Item 12. Operationally, this means the ordinarily untestable assumption that we used to equate choice forms may be untrue for Item 11. A more extensive experiment seems in order.

Note that the differences observed in the trace lines for Item 11, although not quite achieving nominal levels of statistical significance, suggest that Item 11 is

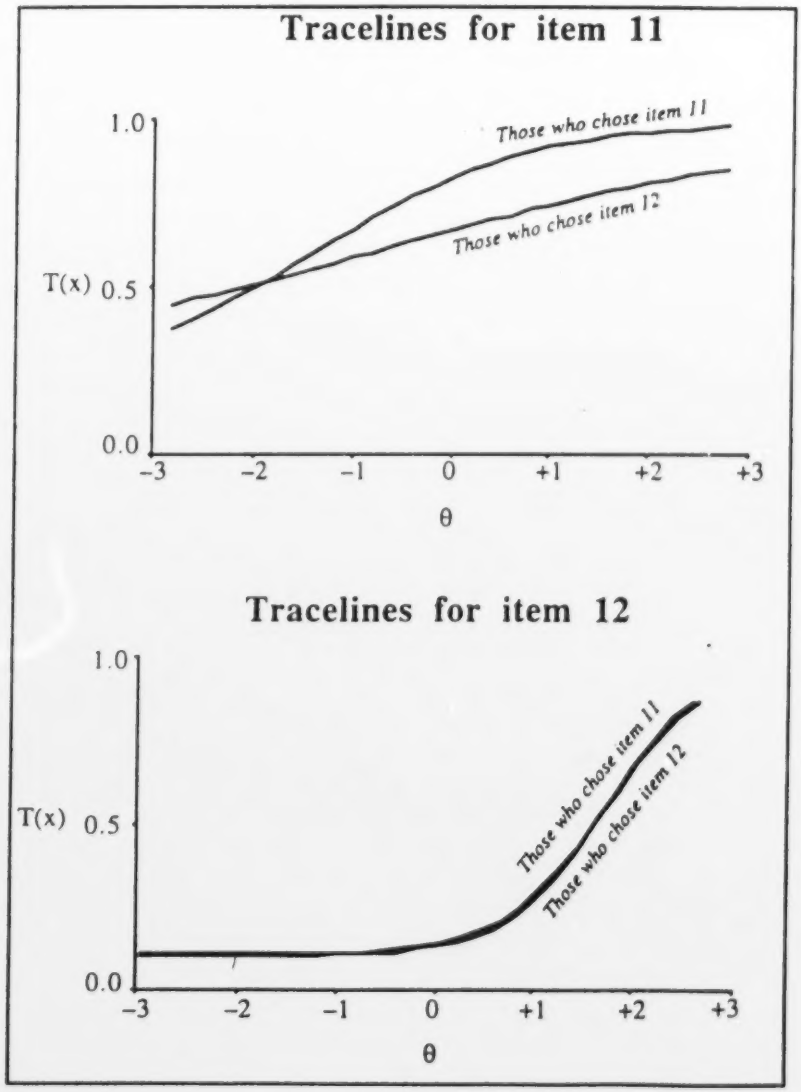

FIGURE 5. *Graphical tests of the ordinarily untestable assumption that choice items have the same trace lines for those who chose them as for those who did not*

easier for those who chose it than for those examinees who did not. This is not always the case. As part of the same study, we found that for another pair of choice items the reverse was true. In none of the cases examined were the differences between $f_1(y)$ and $f_0(y)$ so large as to generate errors in the equating larger than would have been the case had we not equated.

Design 2: Between subjects. In a randomly chosen subset of the examinee population, examinees must be randomly assigned to each of the choice items. This will provide us with unbiased estimates of $g(y)$ for each of the choice items and allow us to equate. It will not provide direct estimates of $f_0(y)$ and $f_1(y)$, but those can be obtained from the portion of the exam in which choice is allowed. As of this writing, an experiment that will have this format is currently being considered for the GRE Writing Test.

Test Dimensionality

Because of the increasing interest in the development of tests that combine the psychometric advantages of multiple-choice items with other features of constructed response items, the following two questions assume importance:

1. Are we measuring the same thing with the constructed response items that we are measuring with the multiple-choice questions?
2. Is it meaningful to combine the scores on the constructed response sections with the multiple-choice score to yield a single reported total score?

Answers to these questions are necessary to build appropriate score-reporting strategies for such hybrid tests. As we shall see, answering these questions is more difficult when the examinee is permitted to choose to answer a subset of the time-consuming constructed response questions (Wainer, Wang, & Thissen, 1991). The use of item response theory to score the test, or to equate forms comprising chosen questions, explicitly requires that the test (or forms) be essentially unidimensional—that all the items measure more or less the same thing. Thus, we must answer the dimensionality question to be able to score the test in a meaningful way. This is explicitly true when using IRT but also must be true when a test score is calculated in many other, less principled rubrics.

Are hybrid tests unidimensional? The literature on this subject is equivocal. Bennett, Rock, Braun, Frye, Spohrer, and Soloway (1991) fitted different factor structures to two relatively similar combinations of multiple-choice, constructed response, and constrained constructed response items; a one-factor model was sufficient for one set of data, but a two-factor model was required for another similar set of data. Bennett, Rock, and Wang (1991) examined a particular two-factor model for the combined multiple-choice and constructed response items on the College Board's Advanced Placement (AP) Test in Computer Science and concluded that the one-factor model provided a more parsimonious fit.

We reanalyzed (Thissen, Wainer, & Wang, 1993) the Computer Science AP data reported by Bennett et al. (1991) and showed that significant, albeit relatively small, factors explain some of the observed local dependence among the constructed response items. We replicated this finding using data from the AP test in chemistry. There was clear evidence that the constructed response problems on both of these tests measure something different than the multiple-choice sections of those tests: There were statistically significant factors for the

constructed response items, orthogonal to the general factor. However, there was also clear evidence that the constructed response problems predominantly measure the same thing as the multiple-choice sections: The factor loadings for the constructed response items were almost always larger on the general (multiple-choice) factor than on the constructed response factor(s). The loadings of the constructed response items on the specifically constructed response factors were small, indicating that the constructed response items do not measure something different very well. Given the small size of the constructed response factor loadings, it is clear that it would take many constructed response items to produce a reliable score on the factor underlying the constructed response items alone—many more items than are currently used.

When we asked the practical question, "Is it meaningful to combine the scores on the constructed response sections with the multiple-choice score to yield a single reported score?" we were driven to conclude that it probably is; indeed, given the small size of the loadings of the constructed response items on their own specific factors, it would probably not be meaningful to attempt to report a constructed response score separately, because it would not be reliably distinct from the multiple-choice score.

Our investigation, and hence the above conclusions, utilized much of the same factor analytic technology, founded on complete data, that has become the standard in dimensionality studies (Jöreskog & Sörbom, 1986, 1988). The procedure assumes that estimates of the covariances were obtained from what is essentially a random sample from the examinee population. However, when there are choice items, assuming a noninformative sampling process[5] is not credible. What is analogous to Assumption 2 that will allow us to factor analyze the observed covariances and treat the results as if they came from the unselected population? Obviously, missing-completely-at-random would suffice, but this is usually patently false in a choice situation.

Can we weaken it? Unfortunately, not much. Suppose we make the obvious assumption that the covariances that we observe are the same as those we do not. Does this allow us to analyze what are observed as if they were the unconditioned covariances? It does not, even with this strong an assumption. To understand why, it is best if we trace the logic mathematically.

What must we assume to allow us to treat $\mathrm{Cov}(y_i, y_j | R_i \times R_j = 1)$ as if they were $\mathrm{Cov}(y_i, y_j)$? There are many possible assumptions. One, parallel to Assumption 2, would be to assume that the covariance involving a choice item is the same among those examinees who did not choose that item as it was among those that did—that is,

$$\text{Condition 1: } \mathrm{Cov}(y_i, y_j | R_i \times R_j = 1) = \mathrm{Cov}(y_i, y_j | R_i \times R_j = 0).$$

But this is not enough. We must also assume that the means for at least one of the two items in the covariance must be the same for those who chose it as it would have been for those who did not.

$$\text{Condition 2: } \mathrm{E}(y_i | R_i = 1) = \mathrm{E}(y_i | R_i = 0) \quad \text{or}$$
$$\mathrm{E}(y_j | R_j = 1) = \mathrm{E}(y_j | R_j = 0).$$

A little algebra will confirm that these conditions will yield the desired result.[6]

How plausible is it that these two conditions will be upheld in practice? Clearly, if one thought that they were likely to be true, what would be the point of providing choice to examinees? Yet, to be able to justify the typical analyses used to answer the crucial dimensionality question, one must posit performance for examinees on the choice items that is essentially the same regardless of whether or not the items were chosen. We find this compelling evidence to look elsewhere for methodologies to answer dimensionality questions when there is choice.

The missing data theory described above presents a convincing argument for the necessity of a special data gathering effort to estimate the covariances associated with choice items. We have demonstrated that there is no easy and obvious model that would allow the credible use of the observed covariances as a proxy for the covariances of interest. To obtain these, we need a special data gathering effort analogous to the ones described earlier. Both kinds of designs require a sample from the unselected population. Design 1 is exactly the same as described earlier. Design 2 is slightly different.

Design 1: Within subjects. In a randomly chosen subset of the examinee population, examinees must be required to indicate their choice but then required to answer all items. As before, this provides estimates of the covariances involving the choice items that are uncontaminated by self-selection. They might suffer the same shortcoming as before; that is, examinees might not be particularly judicious in their choices when they know that they will have to answer all the questions anyway. This will affect any measured relations of y_i and R_i but will probably be satisfactory for estimates of the covariances between the items. We have no data to shed light on these conjectures.

Design 2: Between subjects. In a randomly chosen subset of the examinee population, examinees must be randomly assigned to all pairs of the choice items. This will provide us with unbiased estimates of $\mathrm{Cov}(y_i, y_j)$ for all pairs of the choice items. It will provide more stable estimates of the covariances between each choice item and all of the required items as well. It will thus allow us to do dimensionality studies. Obviously, because this design does not gather any choice information, it cannot provide estimates of $\mathrm{Cov}(y_i, R_i)$.

What Can We Learn From Choice Behavior?

Thus far, our proposed requirements prior to implementing examinee choice fairly require a good deal of work on the part of both the examinee and the examiner. We are aware that extra work and expense are not part of the plan for many choice tests. Often, choice is allowed because there are too many plausible items to be asked and too little time to answer them. Is all of this work really necessary? Almost surely. At a minimum, one cannot know whether it is necessary unless it is done. To paraphrase Derek Bok's comment on the cost of education, if you think doing it right is expensive, try doing it wrong. Yet many well-meaning and otherwise clear-thinking individuals ardently support choice in exams. Why?

The answer to this question must, perforce, be impressionistic. We have heard a variety of reasons. Some are nonscientific; an example is "To show the examinees that we care." The implication is that, by allowing choice, we are giving examinees the opportunity to do their best. We find this justification

difficult to accept, because there is overwhelming evidence to indicate that this goal is unlikely to be accomplished. Which is more important—fairness or the appearance of fairness? Ordinarily, the two go together, but, when they do not, we must be fair and do our best to explain why.

A second justification (W. B. Schrader, personal communication, March 7th, 1993) is that outstanding individuals are usually outstanding on a small number of things. If the purpose of the exam is to find outstanding individuals, we ought to allow them to have the option to show their maximum performance. We find this argument more convincing, but it is moot in a measurement task that is essentially unidimensional.

A third justification might be termed instructional driven measurement (IDM). The argument is that because, in the classroom, students are often provided with choice options evaluation instruments ought to as well. This argument can be compelling, especially if one thinks of the choice options being those that teachers make: which topics to cover, in what order, from what perspective. Why should students suffer the consequences of unfortunate choice that were made on their behalf? The central question is: Can these issues be fairly addressed through the mechanism of allowing choice on exams?

Let us consider more narrowly what we can learn from the choice behavior. Suppose we administer a test that is constructed of two sections. One section is mandatory, and everyone is required to answer all items. A second section contains choice. Equating of different test forms constructed by choice behavior can be done, if we make the usual assumptions required for IRT as well as an assumption about the shape of the choice items' trace lines among those who opted for other items. Suppose, instead, we examine the estimates of proficiency obtained from the mandatory section of the test. How well is proficiency predicted from the choices that examinees make?

An illustration of such a test uses data drawn from the 1989 Advanced Placement Examination in Chemistry (Wainer & Thissen, 1993b). A full description of this test, the examinee population, and the scoring model is found in Wainer, Wang, and Thissen (1991). For the purposes of this illustration, we consider only the five constructed response items in Part II, Section D. Section D has five problems (Problems 5, 6, 7, 8, and 9), of which the examinee must answer three. This section accounts for 19% of the total grade.

Because examinees had to answer three out of the five questions, a total of 10 choice groups was formed, with each group taking a somewhat different test form than the others. Each group had at least one problem in common with every other group; this overlap can be used to place all examinee selected forms on a common scale. The common items serve the role of the mandatory section described earlier. The fitting of a polytomous IRT model to all 10 forms simultaneously was described in Wainer, Wang, and Thissen (1991). As part of this procedure, we obtained estimates of the mean value of each choice group's proficiency (μ_i) as well as the marginal reliability of this section of the test. Our findings are summarized in Table 6. The proficiency scale had a standard deviation of one; those examinees who chose the first three items (5, 6, and 7) were considerably less proficient, on the average, than any other group. The groups labeled 2 through 7 were essentially indistinguishable in performance from one another. Groups 8, 9, and 10 were the best performing groups.

TABLE 6
Summary statistics for the 10 groups formed by examinee choice on Problems 5–9

Group	Problems chosen	Mean group proficiency (μ_i)	n	Cronbach's α
1	5,6,7	−1.02	2,555	0.63
2	6,7,9	−0.04	121	0.65
3	5,6,8	0.00*	5,227	0.57
4	5,7,9	0.04	753	0.64
5	5,7,8	0.08	4,918	0.51
6	6,7,8	0.08	1,392	0.54
7	5,6,9	0.09	457	0.67
8	6,8,9	0.40	407	0.57
9	7,8,9	0.43	898	0.59
10	5,8,9	0.47	1,707	0.59

* The mean for Group 3, the largest group, is fixed at 0.0 to set the location of the proficiency scale.

If we think of Section D as a single item with an examinee falling into one of 10 possible categories, then the estimated proficiency of each examinee is the mean score of everyone in that category. How reliable is this one-item test? We can derive an analog of reliability (see the appendix for a derivation), the squared correlation of proficiency (θ) with estimated proficiency ($\hat{\theta}$), from the between-group variance [var(μ_i)] and the within-group variance (unity). This index of reliability,

$$r^2(\hat{\theta}, \theta) = \text{var}(\mu_i)/[\text{var}(\mu_i) + 1],$$

is easily calculated. The variance of the μ_i is .17, and so $r^2(\hat{\theta}, \theta)$ is .15(= .17/1.17).

It is informative to consider how close .15 is to .57, the reliability of these items when actually scored. Suppose we think of the task of selecting three out of five questions to answer as a single testlet. We can calculate the reliability of a test made up of any number of such testlets using the Spearman-Brown prophesy formula. Thus, if we ask the examinee to pick three from five on one set of topics and then three from five on another, we have effectively doubled the test's length, and its reliability rises from .15 to .26. The estimated reliabilities for tests built of various numbers of such choice testlets are shown in Table 7.

How much information is obtained by requiring examinees to actually answer questions and then grading them? The marginal gain for the AP Chemistry Test is very small; see Figure 6, which shows that at all of the important choice points the error of measurement is virtually the same whether the questions chosen are scored for the content of the answers or scored by noting which choices were made. Thus, we have seen that, for one test, the marginal gain in information by merely noting the choice is almost the same as that which is available from scoring the items. Interestingly, we did not need to make any assumptions about choice behavior, as we did when we scored the item content in the presence of choice-induced missing data, because there is no missing data if the data are the choices.

TABLE 7
Spearman-Brown extrapolation for building a test of specified reliability

Number of testlets*	Reliability
1	0.15
2	0.26
3	0.35
4	0.41
5	0.47
10	0.64
20	0.78

*Here, each testlet comprises the task of selecting three questions out of five.

There is no doubt that more information is available from scoring the constructed response items of the chemistry test than from merely observing which items were chosen to answer. This is reflected in the difference in the size of the reliabilities of the choice test versus the traditionally scored version. This advantage may be diminished considerably on tests based on constructed response items that are holistically scored. Such tests typically have much lower reliability than analytically scored tests.

The reliabilities for the constructed response sections of 20 Advanced Placement Tests are shown in Table 8. Note that there is very little overlap between the

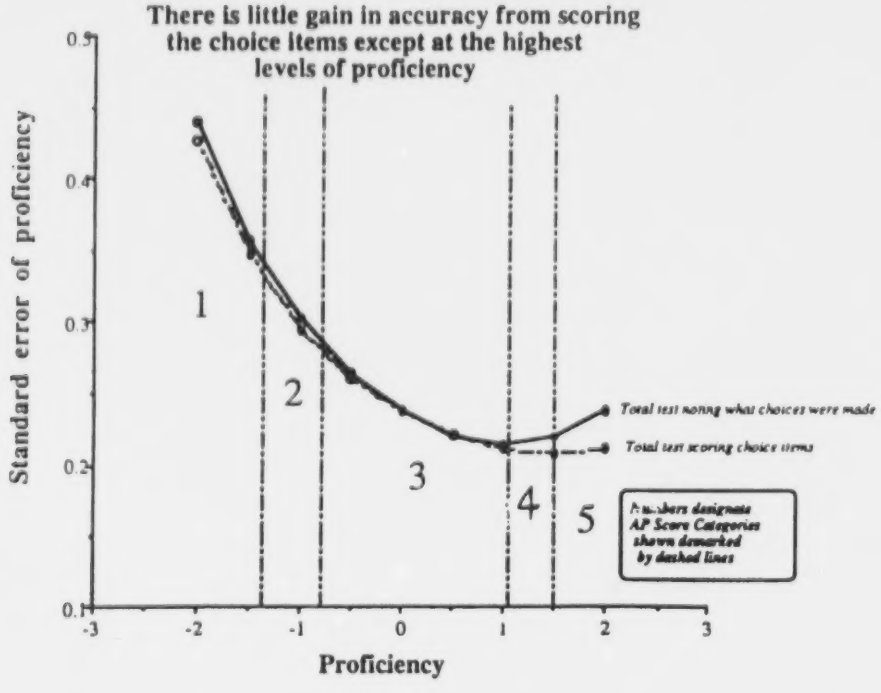

FIGURE 6. *A comparison of the standard errors of estimate of proficiency for two versions of the chemistry test derived by scoring the choice items (84) or merely noting which items were chosen (79). At the selection points of interest, scoring the choice items provides almost no practical increase in precision.*

BEST COPY AVAILABLE

TABLE 8
Reliabilities of constructed response sections of AP tests

Analytically scored	Score reliability	Holistically scored
Calculus AB	0.85	
Physics B	0.84	
Computer Science	0.82	
Calculus BC	0.80	
French Language	0.79	
Chemistry	0.78	
Latin—Virgil	0.77	
Latin—Catullus-Horace	0.76	
Physics C—Electricity	0.74	
Music Theory, Biology	0.73	
Spanish Language	0.72	
Physics C—Mechanics	0.70	History of Art
	0.69	French Literature
	0.63	Spanish Literature
	0.60	English Language & Composition
	0.56	English Literature & Composition
	0.49	American History
	0.48	European History
	0.29	Music: Listening & Literature

distributions of reliability for analytically and holistically scored tests, the latter being considerably less reliable. Chemistry is a little better than average, among analytically scored tests, with a reliability of .78 for its constructed response sections.

It is sobering to consider how well a test that uses only the information about which options are chosen would compare to one of the less reliably scored tests (i.e., any of the holistically scored tests). The structure of such a choice test might be to offer three or four sets of, say, five candidate essay topics, ask the examinees to choose three of those topics in each set that they would write on, and then stop.

Perhaps a more informative analysis of the information available in choices compares it to other sorts of categorical information. Figure 7 shows that more Fisherian information is obtained from examinee choice than is obtained from knowledge of examinee sex and ethnicity but that it is still less than the information obtained from just two (good) multiple-choice items.

It is not our intention to suggest that it is better to have examinees choose questions to answer than it is to actually have them answer them.[7] We observe only that, if the purpose is accurate measurement, some information can be obtained from the choices and that we can obtain this information without relying on untestable (and perhaps unlikely) assumptions about unobservable choice behavior. Moreover, one should feel cautioned if the test administration and scoring scheme yield a measuring instrument little different in accuracy than would have been obtained by ignoring the performance of the examinee entirely.

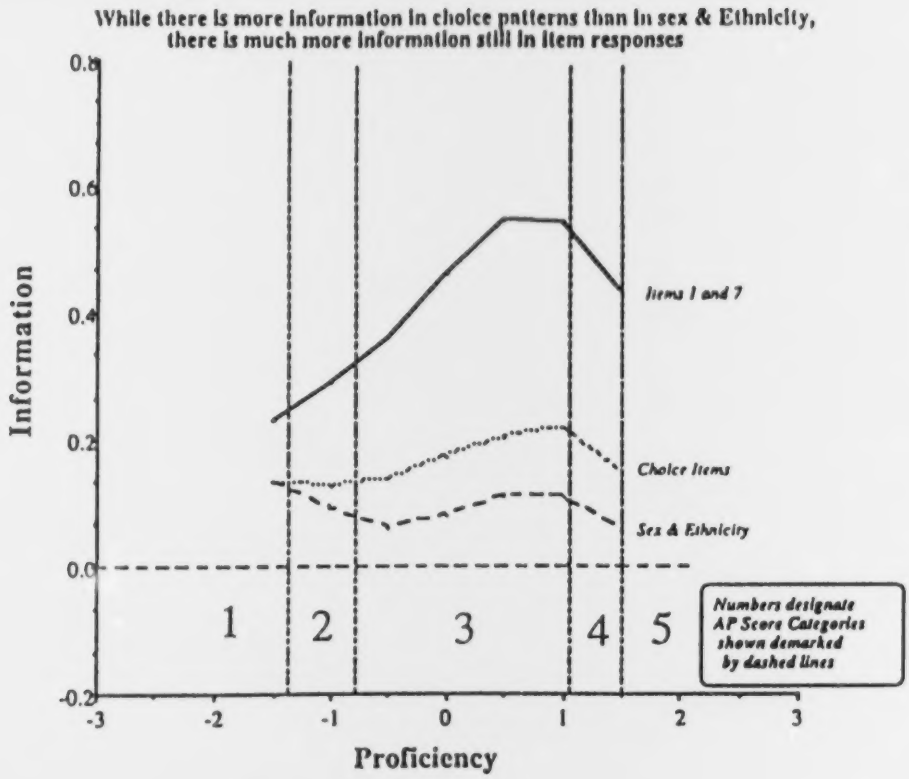

FIGURE 7. *On the AP Chemistry exam, the information about chemistry knowledge provided by just two multiple-choice items dwarfs that available from sex and ethnicity or even choice behavior*

Discussion

We have painted a bleak psychometric picture for the use of examinee choice within fair tests. To make tests with choice fair requires equating the test forms generated by the choice for their differential difficulty. Accomplishing this requires either some special data gathering effort or trust in assumptions about the unobserved responses that, if true, obviate the need for choice. If we can successfully equate choice items, we have thus removed the value of choice in any but the most superficial sense.

To extend these considerations, we need to be explicit about the goals of the test. There are many possible goals of a testing program. In this exposition, we will consider only three: contest, measurement, and device to induce social change.

When a test is a contest, we are using it to determine a winner. We might wish to choose a subset of examinees for admission, for an award, or for a promotion. In a contest, we are principally concerned with fairness. All competitors must be judged under the same rules and under the same conditions. We are not concerned with accuracy, except to require that the test is sufficiently accurate to tell us the order of finish unambiguously.

When a test is used for measurement, we wish to make the most accurate possible determination of some characteristic of an examinee. Usually measurement has some action associated with it; we measure blood pressure and then consider exercise and diet; we measure a child's reading proficiency and then choose suitable books; we measure mathematical proficiency and then choose the next step of instruction. Similarly, we employ measurement to determine the efficacy of various interventions. How much did the diet lower blood pressure? How much better was one reading program than another? When measuring, we are primarily concerned with accuracy. Anything that reduces error may fairly be included on the test.

When a test is a device to induce social change, we are using the test to influence behavior (Torrance, 1993). Sometimes the test is used as a carrot or a stick to influence the behavior of students; we give the test to get students to study more assiduously. Sometimes the test is used to influence the behavior of teachers; we construct the test to influence teachers' choice of material to be covered. The recent literature (Popham, 1987) has characterized this goal as measurement driven instruction (MDI). MDI has engendered rich and contentious discussions, and we will not add to them here. The interested reader can begin with Cizek (1993) and work backward through the references provided by him. At first, it might appear that, when a test is being used in this way, issues of fairness and measurement precision are not important, although the appearance of fairness may be. However, that is false. When a test is used to induce change, the obvious next question must be, "How well did it work?" If we used the test to get students to study more assiduously, or to study certain specific material, or to study in a different way, how much did they do so? How much more have the students learned than they would have under some other condition? The other condition might be no announced test, or it might be with a test of a different format. There are obvious experimental designs that would allow us to investigate such questions—but all require measurement.[8] Thus, even when the purpose of the test is to influence behavior, that test still ought to satisfy the canons of good measurement practice.

Thus far, we have confined our discussion to situations in which it is reasonable to assign any of the choice items to any examinee. Such an assumption underlies the notions of equating, which as we have used the term requires essential unidimensionality (using Stout's, 1990, useful terminology), and also of the experiments we have described to ascertain the difficulty of the choice items in the unselected population. Situations in which examinees are given such a choice we call *small choice*.

Small choice is used most commonly because it is felt that measurement of the underlying construct may be contaminated by the particular context in which the material is embedded. It is sometimes thought that by allowing examinee choice from among several different contexts a purer estimate of the underlying construct may be obtained. Consider, for example, the following two math problems that are intended to test the same conceptual knowledge:

1. The distance between the Earth and the Sun is 93 million miles. If a rocket ship took 40 days to make the trip, what was its average speed?
2. The Kentucky Derby is one and one-fourth miles in length. When Northern Dancer won the race with a time of 2 minutes, what was his average speed?

The answer to both problems may be expressed in miles/hour. Both problems are formally identical, except for differences in the difficulty of the arithmetic. Allowing an examinee to choose between these items might allow us to test the construct of interest (Does the student know the relation Rate × Time = Distance?), while at the same time letting the examinees pick the context within which they feel more comfortable.

Big choice. In contrast to small choice is a situation in which it makes no sense to insist that all individuals attempt all tasks (e.g., it is of no interest or value to ask the editor of the school yearbook to quarterback the football team for a series of plays in order to gauge proficiency in that context). We call this sort of situation *big choice*. Using more precise language, we would characterize situations involving big choice as multidimensional. Making comparisons among individuals after those individuals have made a big choice is quite common. College admissions officers compare students who have chosen to take the French Achievement Test against those who opted for one in physics, even though their scores are on completely different scales. Companies that reward employees with merit raises usually have a limited pool of money available for raises and, in the quest for an equitable distribution of that pool, must confront such imponderable questions as "is person A a more worthy carpenter than person B is a statistician?" At the beginning of this account, we set aside big choice while we attempted to deal with the easier problems associated with small choice.

Most of what we have discussed so far leans heavily on sampling responses in an unselected population and thus applies primarily to the small choice situation. Can we make useful comparisons in the context of big choice? Yes, but only for tests as contests, at least for the moment. When there is big choice, we can set out rules that will make the contest fair. We are not able to make the inferences that are usually desirable for measurement. To illustrate, let us consider the scoring rules for the decathlon as an illustration of scoring a multidimensional test without choice. Building on this example, we will expand to the situation of multidimensionality and choice.

The decathlon is a 10-part track event that is clearly multidimensional. There are strength events like discus, speed events like the 100m dash, endurance events like the 1,500m run, and events that stress agility, like the pole vault. Of course, underlying all of these events is some notion of generalized athletic ability, which may predict performance in all events reasonably accurately.[9] How is the decathlon scored? In a word, arbitrarily. Each event is counted "equally" in that an equal number of points is allocated for someone who equaled the world record that existed in that event at the time that the scoring rules were specified.[10] How closely one approaches the world record determines the number of points received (i.e., if one is within 90% of the world record, one gets 90% of the points). As the world record in separate events changes, so too does the number of points allocated. If the world record got 10% better, then 10% more points would be allocated to that event. Let us examine the two relevant questions: Is this accurate measurement? Is this a fair contest?

To judge the accuracy of the procedure as measurement, we need to know the qualities of the scale so defined. Can we consider decathlon scores to be on a ratio scale? Is an athlete who scores 8,000 points twice as good as someone who scores

4,000? Most experts would agree that such statements are nonsensical. Can we consider decathlon scores to be on an interval scale? Is the difference between an athlete who scores 8,000 and one who scores 7,000 in any way the same as the difference between one who scores 2,000 and another who scores 1,000? Again, experts agree that this is not true in any meaningful sense.

Can we consider decathlon scores to be ordinally scaled? Yes. A demonstration uses standard mathematical notation and is virtually identical to the description given in Krantz, Luce, Suppes, and Tversky (1971, p. 14):

Definition: Let A *be a set and* $\succeq$ *be a binary relation on* A, *i.e.* $\succeq$ *is a subset of* $A \times A$. *The relational structure* $(A, \succeq)$ *is a weak order if and only if, for all* $a, b, c \in A$, *the following two axioms are satisfied:*

1. *Connectedness: Either* $a \succeq b$ *or* $b \succeq a$.
2. *Transitivity: If* $a \succeq b$ *and* $b \succeq c$, *then* $a \succeq c$.

If such a definition holds, it can be proved that

If A *is a finite nonempty set and if* $(A, \succeq)$ *is a weak order, then there exists a real-valued function* ϕ *on* A *such that for all* $a, b \in A$,

$a \succeq b$ *if and only if* $\phi(a) \geq \phi(b)$.

ϕ *is then an ordinal scale.*

Translating this into the current context, *A* might represent the collection of performances on one of the various decathlon events, scaled in seconds or meters or whatever. ϕ is the scoring function that translates all of those performances into points. It is straightforward to examine any particular scoring function to see if it satisfies these conditions. Obviously, any function that is monotonic will satisfy them.

We conclude that decathlon scoring satisfies the conditions for an ordinal scale. A fair contest must. This raises an important and interesting issue: If we are using a test as a contest and we wish it to be fair, we must gather data that would allow us to test the viability of the assumptions stated in the definition above. The most interesting condition is that of transitivity. The condition suggests two possible outcomes in a situation involving multidimensional comparisons:

1. There may exist instances in which Person A is preferred to Person B and Person B to Person C, and, last, Person C is preferred to Person A. This happens sufficiently often so that we cannot always attribute it to random error. It means that, in some multidimensional situations, no ordinal scale exists.
2. Data that allow the occurrence of an intransitive triad are not gathered. This means that while the scaling scheme may fail to satisfy the requirements of an ordinal scale, which are crucial for a fair contest, we will never know.

In a situation involving big choice, we do not know if the connectedness axiom is satisfied. How can we test the viability of this axiom if we can observe only *a* on one person and only *b* on another?

To get a better sense of the quality of measurement represented by the decathlon, let us consider what noncontest uses might be made of the scores. The most obvious use would be as a measure of the relative advantage of different training methods. Suppose we had two competing training methods—for example, one emphasizing strength and the other endurance. We could then conduct an experiment in which we randomly assigned athletes to one or the other of these two methods. In a pretest, we could get a decathlon score for each competitor and then another after the training period had ended. We could then rate each method's efficacy as a function of the mean improvement in total decathlon score. While one might find this an acceptable scheme, it may be less than desirable. Unless all events showed the same direction of effect, some athletes might profit more from a training regime that emphasizes strength; others might need more endurance. It seems that it would be far better not to combine scores but, instead, to treat the 10 component scores as a vector. Of course, each competitor would almost surely want to combine scores to see how much his total had increased, but that is later in the process. The measurement task, from which we are trying to understand the relation between training and performance, is better done at the disaggregated level. It is only for the contest portion that the combination takes place.

We conclude that scoring methods that resemble those used in the decathlon can only be characterized as measurement in an ordinal sense. And thus, the measures obtained are only suitable for crude sorts of inferences.

When Is a Contest Fair?

In addition to the requirement of an ordinal scale, fair measurement also requires that all competitors know the rules in advance, that the same rules must apply to all competitors equally, and that there is nothing in the rules that gives one competitor an advantage over another because of some characteristic unrelated to the competition. How well do the decathlon rules satisfy these criteria?

Certainly the scoring rules, arcane as they might be, are well known to all competitors, and they apply evenhandedly to everyone. Moreover, the measurements in each event are equally accurate for every competitor. Thus, if two competitors both throw the shot the same distance, they will get the same number of points. Last, is a competitor placed at a disadvantage because of unrelated characteristics? No; each competitor's score is determined solely by his performance in the events. We conclude that the decathlon's scoring rules comprise a fair contest even though they comprise a somewhat limited measuring instrument.

The decathlon represents a good illustration of what can be done with multidimensional tests. Sensible scoring can yield a fair contest, but it is not good measurement. There has been an attempt to somehow count all events equally, balancing the relative value of an extra inch in the long jump against an extra second in the 1,500 meter run. But no one would contend that they are matched in any formal way. Such formal matching is possible, but it requires agreement on the metric. The decathlon is a multidimensional test, but it is not big choice as we have previously defined it. Every competitor provides a score in each event (on every item). How much deterioration would result if we add big choice into this mix?

Big choice makes the situation worse. One may be able to invent scoring rules that yield a fair contest but do not give an accurate measurement. As one example, consider ABC's "Super Star's Competition," a popular TV pseudosport in which athletes from various sports are gathered together to compete in a series of seven different events. The athletes each select five events from among the seven. The winner of each event is awarded 10 points, second place 7, third place 5, and so on. The overall winner is the one who accumulates the most points. Some events are "easier" than others because fewer and/or lesser athletes elected to compete in that event; nevertheless, the same number of points are awarded. This is big choice by our definition, in that there are events that some athletes could not compete in (i.e., Joe Frazier, a former world champion boxer, chose not to compete in swimming because he could not swim). Are the scores in such a competition measurement? No. Is the contest fair? By the rules of fairness described above, yes, although the missingness of some of the data makes checking key underlying assumptions problematic.

The current state of the art allows us to use big choice in a multidimensional context and, under limited circumstances, to have fair contests. We cannot yet have measurement in this context at a level of accuracy that can be called anything other than crude. As such, we do not believe that inferences based on such procedures should depend on any characteristic other than their fairness. This being the case, users of big choice should work hard to assure that their scoring schemes are indeed as fair as they can make them. Wainer (1993) and Wainer and Deveaux (1994) provide two detailed case studies describing how this might be accomplished.

When is it not fair? Paul Holland (Allen, Holland, & Thayer, 1993, p. 5) calls big choice "easy choice," because often big choice is really no choice at all. Consider a choice item in which an examinee is asked to discuss the plot of either (a) *The Pickwick Papers* or (b) *Crime and Punishment* from a Marxist perspective. If the student's teacher chose *The Pickwick Papers*, there really is no choice. At least, the student had no choice. Because many times in a big choice situation the examinee really has no choice, in that it is not plausible to answer any but a single option, fairness requires the various options to be of equal difficulty. This returns us to the primary point of this account. How are we to ascertain the relative difficulty of big choice items?

Is Big Choice Useful When the Test's Goal Is to Induce Social Change?

If we wish to use the test to influence instruction, we might evaluate the success of the enterprise by surveying the field before and after the test became widespread. But this is surely only a superficial goal. The primary goal is not the structure of instruction but rather the effects of that instruction on the students. Thus, any attempt to measure the efficacy of an intervention (in this case a particular kind of test structure) must eventually use some sort of measuring instrument. We must also pay careful attention that the use of a test to induce change does not compromise its fairness. We know of one standardized science test that introduced a very easy item on a new topic as a possible choice. The goal was to influence teachers to cover this new area. Examinees whose teachers covered this topic had a distinct advantage over examinees whose teachers had not. Since the choice was really made months before the test, and by the teacher, not the student, is this fair?

Conclusions

This summary of research is far from conclusive; many questions remain. It would be good to know how far away from unidimensionality a test can be and still yield acceptable measurement when choices are allowed. How far from ignorable can nonresponse be and still be acceptably adjusted for statistically? What kinds of conditioning variables are helpful in such adjustments? What are the most efficient kinds of data-gathering designs? Such questions lend themselves to solutions through careful experimentation and computer simulation.

Can the uncritical use of choice lead us seriously astray? While there are several sources of evidence summarized in this article about the size of choice effects, we focused on just one series of exams. Summaries from other sources, albeit analyzed in several different ways, lead us to believe that the Advanced Placement Tests, often referred to because they currently involve choice, are not unusual. In fact, they may be considerably better than average. A recent experience with allowing choice in an experimental SAT is instructive (Lawrence, 1992). It has long been felt by math teachers that it would be better if examinees were allowed to use calculators on the mathematics portion of the SAT. An experiment was performed in which examinees were allowed to use a calculator, if they wished. The hope was that it would have no effect on the scores. Calculators did improve scores. The experiment also showed that examinees who used more elaborate calculators got higher scores than those who used more rudimentary ones. Sadly, a preliminary announcement had already been made indicating that the future SAT-M would allow examinees the option of using whatever calculator they wished, or not using one at all.

A testing situation corresponds to measuring people's heights by having them stand with their backs to a wall. Allowing examinees to bring a calculator to the testing situation, or not, but not knowing for sure whether they had one, or what kind, corresponds to having some persons to be measured for height, unbeknownst to you, bring a stool of unknown and varying height on which to stand. Accurate and fair measurement is no longer possible in either case.

Our discussion has concentrated on explicitly defined choice in tests, or *alternative questions* in the language of the first half of this century. However, in the case of portfolio assessment, the element of choice is implicit and not amenable to many of the kinds of analysis that have been described here. Portfolio assessment may be more or less structured in its demands on the examinee—that is, it may specify the elements of the portfolio more or less specifically. However, to the extent that the elements of the portfolio are left to the choice of the examinee, portfolio assessment more closely resembles ABC's "Super Star's Competition" than even the decathlon. In portfolio assessment, how many forms of the test are created by examinee choice? Often, as many as there are examinees! If that is the case, can those forms be statistically equated? No. This fact has clear consequences in the results obtained with portfolio assessment; for instance, Koretz, McCaffrey, Klein, Bell, and Stecher (1992) report that the reliability of the 1992 Vermont portfolio program measures was substantially less than is expected for useful measurement. Can it be otherwise, when the examinees (effectively) construct their own tests?[11]

Is building examinee choice into a test possible? Yes, but it requires extra work. Approaches that ignore the empirical possibility that different items do

not have the same difficulty will not satisfy the canons of good testing practice, nor will they yield fair tests. But, to assess the difficulty of choice items, one must have responses from an unselected sample of fully motivated examinees. This requires a special sort of data gathering effort.

What are we estimating when we use examinee selected items? If we are interested in θ_{Max}, then we need to choose the items for the examinees. The belief that the estimate of θ_{Max} obtained from examinee selected items is accurate has been disconfirmed by the data gathered so far. Although these data are of modest scope, they indicate what sorts of data need to be gathered to examine this question more fully.

What can we do if the assumptions required for equating are not satisfied across the choice items? If test forms are built that cannot be equated (made comparable), scores comparing individuals on incomparable forms have their validity compromised by the portion of the test that is not comparable. Thus, we cannot fairly allow choice if the process of choosing cannot be adjusted away.

Choice is anathema to standardized testing unless those aspects that characterize the choice are irrelevant to what is being tested.

Notes

[1] Section II of the 1921 exam asked the examinee to answer 5 of 26 questions. This alone yielded more than 65 thousand different possible "forms." When coupled with Section III (pick one essay topic from among 15) and Section I ("Answer 1 of the following 3"), we arrive at the unlikely figure shown in Table 1.

[2] We will use both the language and notation of item response theory (IRT). This is not necessary; our argument could be phrased in traditional true score theory terms. We chose to place this argument within an IRT framework because it allows greater precision of explanation. This is especially important in later sections where being explicit about the estimand and the assumptions is critical.

[3] We are oversimplifying IRT scoring here; for most IRT models, the score associated with each item response actually depends on the other item responses, and so it may be different for each examinee.

[4] This varies a little from Little and Rubin's (1987) conception. They would require the independence of choice given some observed conditioning variable. In our construction, the conditioning variable, θ, is latent. In any operational test, this difference is only a technical one for, when the test is longish, raw score (observable) and θ can be transformed from one to the other easily. This does not apply in situations like adaptive testing in which raw score is unrelated to θ.

[5] A sampling process is noninformative in this case if, by knowing an individual's choice, we learn nothing about how well they will do on the item.

[6] Our thanks to Nick Longford for pointing this out to us.

[7] Our colleague Nick Longford commented that this "suits perfectly the current American culture in which no one ever actually does anything but is concerned instead with management."

[8] It is not uncommon in education for innovations to be tried without an explicit design to aid in determining the efficacy of the intervention. Harold Gulliksen (personal communication, October 26, 1965) was fond of recounting the response he received when he asked what the control condition was against which the particular education innovation was to be measured. The response was "We didn't have a control because it was only an experiment."

[9] Actually, it only predicts accurately for top-ranked competitors, who tend to perform "equally" well in all events. There are some athletes who are very much

better in one event or another, but they tend to have much lower overall performance than generalists who appear more evenly talented.

[10] The Olympic Decathlon scoring rules were first established in 1912 and allocated 1,000 points in each event for a world record performance. These scoring rules have been revised in 1936, 1950, 1964, and 1985. It is interesting to note (Mislevy, 1992) that the 1932 gold medal winner would have finished second under the current (1985) rules.

[11] The idea of portfolio assessment includes two components, one of which is examinee choice of material to submit, and the other is that the material is collected over some longer period of time than in a conventional test. The latter idea, collecting responses over a long period of time, is certainly a useful one. However, the former idea, letting the examinees choose their test, leads to noncomparable (and unreliable) scores. Long-term data collection is certainly possible with well-specified prompts, questions, or items that leave the examinee no choice. Portfolio assessment would provide better measurement to the extent that the element of choice was removed.

[12] We are grateful to Charles Lewis who suggested this analog for reliability, provided a derivation, and cautioned against its too broad usage.

APPENDIX [12]

How can we calculate a reliability coefficient from the classification of examinees by their choice of items? Let us assume that we know the mean proficiency of all examinees in each choice group. We will index examinees by j and choice groups by i, and the model we use is

$$\theta_{ij} = \mu_i + z_{ij} \tag{A1}$$

where the proficiency of person j in group i is θ_{ij} and is distributed normally with mean μ_i and variance 1. We represent the deviation of each person j within group i from that group's mean as z_{ij}.

If we estimate $\hat{\theta}_{ij}$ with μ_i, the mean of group i, that is

$$\hat{\theta}_{ij} = \mu_i. \tag{A2}$$

The correlation between θ_{ij} and $\hat{\theta}_{ij}$ is analogous to validity if we think of θ_{ij} as the analog of true score. The square of this correlation can be thought of as a measure of reliability. Keeping this in mind, we can derive a computational formula for

$$r^2(\hat{\theta}_{ij}, \theta_{ij})$$

by noting that

$$r^2(\hat{\theta}_{ij}, \theta_{ij}) = [\mathrm{cov}(\hat{\theta}_{ij}, \theta_{ij})]^2/[\mathrm{Var}(\theta_{ij}) \times \mathrm{Var}(\hat{\theta}_{ij})]. \tag{A3}$$

In the numerator,

$$\begin{aligned} \mathrm{cov}(\hat{\theta}_{ij}, \theta_{ij}) &= \mathrm{cov}[\mathrm{E}(\theta_{ij} \mid i), \mathrm{E}(\hat{\theta}_{ij} \mid i)] + \mathrm{E}[\mathrm{Cov}(\hat{\theta}_{ij}, \theta_{ij} \mid i)] \\ &= \mathrm{cov}(\mu_i, \mu_i) + \mathrm{E}[\mathrm{Cov}(\theta_{ij}, \mu_i \mid i)] \\ &= \mathrm{var}(\mu_i). \end{aligned}$$

The rightmost term in the initial expression $\{\mathrm{E}[\mathrm{Cov}(\hat{\theta}_{ij}, \theta_{ij} \mid i)]\}$ is zero, and hence the expression reduces to the covariance of μ_i with itself or the variance of μ_i.

This is the expression in the numerator that we need to compute (A3). The denominator requires the variance of both $\hat{\theta}_{ij}$ and θ_{ij}. These are easily computed from

$$\begin{aligned} \mathrm{Var}(\theta_{ij}) &= \mathrm{Var}[\mathrm{E}(\theta_{ij} \mid i)] + \mathrm{E}[\mathrm{Var}(\theta_{ij} \mid i)] \\ &= \mathrm{var}(\mu_i) + 1, \end{aligned} \tag{A4}$$

and

$$\mathrm{Var}(\hat{\theta}_{ij}) = \mathrm{var}(\mu_i). \tag{A5}$$

Substituting these results into (A3) yields

$$r^2(\hat{\theta}_{ij}, \theta_{ij}) = \mathrm{var}(\mu_i)/[\mathrm{var}(\mu_i) + 1]. \tag{A6}$$

The estimate of $\mathrm{var}(\mu_i)$ we obtained from Section D of AP Chemistry is .17, and hence the estimated reliability [from (A6)] is .15.

References

Allen, N. L., & Holland, P. W. (1993). A model for missing information about the group membership of examinees in DIF studies. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 241–252). Hillsdale, NJ: Erlbaum.

Allen, N. L., Holland, P. W., & Thayer, D. T. (1993). The optional essay problem and the hypothesis of equal difficulty (ETS Tech. Rep. No. 93-94). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1991). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151–162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92.

Brigham, C. C. (1934). *The reading of the comprehensive examination in English*. Princeton, NJ: Princeton University Press.

Cizek, G. J. (1993). Rethinking psychometricians' beliefs about learning. *Educational Researcher, 22*(4), 4–9.

College Entrance Examination Board. (1905). *Questions set at the examinations held June 19–24, 1905*. New York: Ginn.

College Entrance Examination Board. (1990). *The 1989 Advanced Placement Examinations in Chemistry and their grading*. Princeton, NJ: Advanced Placement Programs.

DeMauro, G. E. (1991). *The effects of the availability of alternatives and the use of multiple choice or essay anchor tests on constructed response constructs* (Draft Report). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1990). Scaling and equating. In H. Wainer with N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, and D. Thissen, *Computerized adaptive testing: A primer* (pp. 137–160). Hillsdale, NJ: Erlbaum.

Fitzpatrick, A. R., & Yen, W. M. (1993, April). The psychometric characteristics of choice items. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

Fremer, J., Jackson, R., & McPeek, M. (1968). *Review of the psychometric characteristics of the Advanced Placement Tests in Chemistry, American History, and French* (Internal Memorandum). Princeton, NJ: Educational Testing Service.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 115–142). New York: Springer-Verlag.

Gulliksen, H. O. (1950). *A theory of mental tests*. New York: Wiley. (Reprinted, 1987, Hillsdale, NJ: Erlbaum).

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Jöreskog, K. J., & Sörbom, D. (1986). *PRELIS: A program for multivariate data screening and data summarization*. Chicago, IL: Scientific Software.

Jöreskog, K. J., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS.

Kierkegaard, S. (1986). *Either/or*. New York: Harper & Row.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (Interim Report). Santa Monica, CA: RAND Institute on Education and Training.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. 1*. New York: Academic.

Lawrence, I. (1992). *Effect of calculator use on SAT-M score conversions and equating* (Draft Report). Princeton, NJ: Educational Testing Service.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lukhele, R., Thissen, D., & Wainer, H. (1993). *On the relative value of multiple-choice, free-response, and examinee-selected items in two achievement tests* (ETS Tech. Rep. No. 93-28). Princeton, NJ: Educational Testing Service. (Also in press, *Journal of Educational Measurement, 31.*)

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Draft Report). Princeton, NJ: Educational Testing Service.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). *Choice in Advanced Placement Tests* (Unpublished Statistical Report No. SR-92-51). Princeton, NJ: Educational Testing Service.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*, 679–682.

Powers, D. E., Fowles, M. E., Farnum, M., & Gerritz, K. (1992). *Giving a choice of topics on a test of basic writing skills: Does it make any difference* (Research Report No. 92-19)? Princeton, NJ: Educational Testing Service.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thissen, D., Wainer, H., & Wang, X. B. (1993). *How unidimensional are tests comprising both multiple-choice and free-response items? An analysis of two tests* (ETS Tech. Rep. No. 93-32). Princeton, NJ: Educational Testing Service. (Also in press, *Journal of Educational Measurement, 31.*)

Torrance, H. (1993). Combining measurement-driven instruction with authentic assessment: Some initial observations of the national assessment in England and Wales. *Educational Evaluation and Policy Analysis, 15*, 81–90.

Wainer, H. (1993). How much more efficiently can humans run than swim? *Chance, 6*, 17–21.

Wainer, H., & Deveaux, R. (1994). Resizing triathlons for fairness. *Chance*, *7*(1), 20-25.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197–219.

Wainer, H., & Thissen, D. (1993b). *Choosing: A test* (ETS Tech. Rep. No. 92-25). Princeton, NJ: Educational Testing Service.

Wainer, H., Wang, X. B., & Thissen, D. (1991). *How well can we equate test forms that are constructed by examinees* (Tech. Rep. No. 91-15)? Princeton, NJ: Educational Testing Service. (Also in press, *Journal of Educational Measurement*, *31*.)

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*, 373–391.

Wang, X. B. (1992). *Achieving equity in self-selected subsets of test items*. Unpublished doctoral dissertation, University of Hawaii at Manoa, Honolulu.

Wang, X. B., Wainer, H., & Thissen, D. (1993). *On the viability of some untestable assumptions in equating exams that allow examinee choice* (ETS Tech. Rep. No. 93-31). Princeton, NJ: Educational Testing Service.

Authors

HOWARD WAINER is Principal Research Scientist, Educational Testing Service, T-15, 666 Rosedale Rd., Princeton, NJ 08541. He specializes in statistics and psychometrics.

DAVID THISSEN is Professor and Director of the Graduate Program in Quantitative Psychology and Acting Director of the L. L. Thurstone Psychometric Laboratory at the University of North Carolina, Chapel Hill, CB# 3270, Davie Hall, Chapel Hill, NC 27599-3270. He specializes in psychometrics and quantitative psychology.

Received June 25, 1993
Revision received September 20, 1993
Accepted October 4, 1993

BEST COPY AVAILABLE

54020-11885 • Y104P550 • 209025